KARAKTERSZELEKCIÓ JELENTŐSÉGE MORFOMETRIAI VIZSGÁLATOKNÁL

Demeter János
Természettudományi Muzeum Állattára

Bevezetés

A többváltozós statisztikai módszerek egyre szélesebb körü ismerete és a feldolgozást lehetővé tevő számitógépek egyre növekvő száma azt eredményezte, hogy egyre több morfometriai vizsgálat sorakoztatja fel eszköztárában elsősorban a faktoranalizist és a kanonikus analizist [1].

Az adott vizsgálat szempontjából releváns és szükséges számu bélyegek kiválasztása egyik alapfeltétele a megbizható, reprodukálható eredményü elemzésnek. Sokal és Sneath [12], csupán a korrelációs együttható konfidencia intervallumára alapozva véleményüket, a bélyegek minimális számát hatvanra becsülték. Mivel a morfometriai vizsgálatok megfigyelési egységei egyedek, az N x P alapadat mátrixból P x P hasonlósági mátrixot számitunk ki, és a mintanagyság (N) ritkán esik a kritikusnak vélt szint alá.

A megfigyelési egységek n-dimenziós absztrakt morfológiai térben léteznek. Az absztrakt tér dimenzióiból P ($P<N$) számu bélyeget ismerünk, de ezek alapján is tudjuk, hogy a karakterek közötti korreláció-kovariancia folytán ez az absztrakt tér nem ortogonális tengelyek által határolt. A P számu karakter megismerése is gyakorlati akadályba ütközik, ezért a kérdésfeltevés: létezik-e a végtelen számu karaktert jól közelitő P számu bélyeg olyan részhalmaza, amely minimális torzitással, adekvát módon reprezentálja P-t?

A háttérváltozók keresésének egyik klasszikus módszere a faktoranalizis. Davies és Boratynski [3] azt találták, hogy a legnagyobb öt sajátértékhez tartozó sajátvektorok legnagyobb abszolut értékü elemei segitségével az eredeti 101 karaktert huszonötre sikerült redukálni. Egy későbbi dolgozatában Davies [4] továbbfejlesztette korábbi vizsgálatait és a karakterek információtartalma szerinti rangsorolást találta a legalkalmasabbnak a Hemiptera rovarrend egyik családjának numerikus taxonómiai elemzésekor.

Szeretném bemutatni dolgozatomban, hogy az extrahált faktorok értelmezése milyen nehézségekbe ütközhet morfometriai vizsgálatoknál, s ezért a bélyegek információtartalmuk szerinti rangsorolásának, valamint a diszperziós kritérium szerinti rangsorolásnak [7] az alkalmazhatóságát kivánom feltárni.

Anyag és módszer

Az *Apodemus* rágcsálógenusz három fajának 156 példányáról 53 koponyaméretet és 12 külső bélyeget vettünk fel. A bélyegek részletes leirása irreleváns lenne e helyen, a vizsgálat ilyen jellegü részletei másutt [4] megtalálhatóak.

A 65 bélyeg faktoranalizise főkomponens faktor-extrakció segitségével, forgatás nélkül történt a BMDP4M program [5] felhasználásával. A főkomponensanalizis módszere közismert [6], ezért ennek, vala-

mint az információtartalom entrópia-függvény segitségével történő becslésének [11] részletezését mellőzöm.

Orlóci [7] diszperziós kritérium módszerének kiindulása:

$$SS = \max \left[\sum_h S_{h1}^2 / S_{11}, \ldots, \sum_h S_{hP}^2 / S_{PP} \right], \quad h=1,\ldots,P \qquad (1)$$

Tehát az a bélyeg, amelynek részesedése legnagyobb a totális négyzetösszegből, kapja az 1. rangot.

A következő lépésben a reziduális négyzetösszegek és keresztszorzatok kiszámitás történik,

$$S_{hi} := S_{hi} - (S_{hm}/\sqrt{S_{mm}}) \cdot (S_{im}/\sqrt{S_{mm}}), \qquad (2)$$

bármelyik h,i=1,...,P-re.

Az eljárást lépésenként P-szer megismételjük, s ezzel az összes varianciát ortogonális tagokra bontjuk. Az SS értékek felfoghatók mint "saját szórás", és igy a $SS/\sum_h S_{hh}$, h=1,...,P arány a relativ fontosság mértéke. Orlóci [7,8] rámutatott arra, hogy ellentétben a faktoranalizissel, amely a sajátérték-probléma megoldásán keresztül végzi el az ortogonális particiót és a közös szórás-komponens elemzése révén keres háttér-faktorokat, a diszperziós kritérium szerinti rangsorolás a saját szórás feltárására helyezi a hangsulyt.

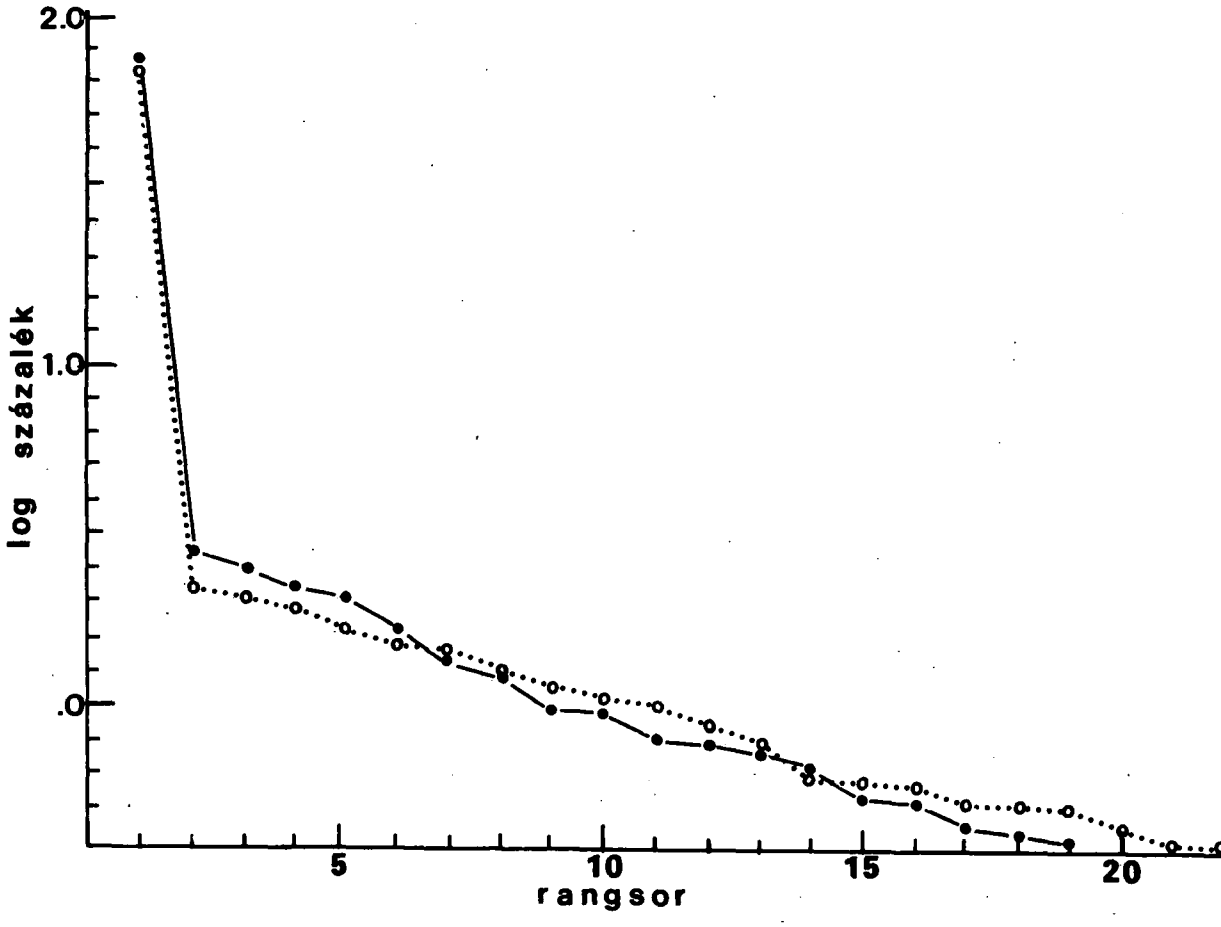
Eredmények

Az 1. ábrán bemutatom, hogy a főkomponensanalizis által létrehozott mesterséges változók közül az első 19 felel az összes variancia 95 %-ért, mig Orlóci módszere szerint 22 változó kumulativ saját szórása felelős 95 %-ban az összvarianciáért. Feltünő, hogy a két görbe lefutása rendkivül hasonló, az első főkomponens, valamint a legnagyobb specifikus szórásu változó, DIM3 (a metszőfog és a harmadik zápfog közti távolság), egyaránt több mint 70 %-ban reprezentálják az összes varianciát. A további komponensek illetve változók már jóval kisebb arányban járulnak hozzá az adatstruktura alakulásához.

A két görbe kvantitativ lefutása arra engedne következtetni, hogy lényegében azonos információt kapunk a két módszerrel. Azonban vizsgáljuk meg az első öt főkomponens faktorsuly-mátrixát /1. tábl./.

Belátható, hogy igen nehéz az uj mesterséges változók értelmezése, mivel az első főkomponens, amely az összvariancia több mint kétharmadát foglalja magába, az eredeti változók tulnyomó részével szoros korrelációt mutat. Az azonban látható, hogy a diszperziós kritérium módszerével legnagyobb specifikus szórásunak talált bélyeg, DIM3, egyike a legnagyobb sulyu változóknak.

A bélyegek információtartalmát kifejező entrópia-függvény (H') használhatósága sulyos korlátokba ütközött, amely részletezésére másutt térek ki [4]: a 2. ábrán látható, hogy a bélyegek információtartalma a karakter-állapotok diverzitásának mértékeként a metrikus karakterek szórásának függvénye, igy a fentemlitett additiv variancia-komponensek miatt belátható, hogy az információtartalom felbontása közös, valamint specifikus komponensekre ugyszintén kivánatos volna [9].



1. ábra. A főkomponensváltozók /folyamatos vonal/ és a diszperziós kritérium szerint rangsorolt eredeti változók saját szórásának /szaggatott vonal/ százalékos részesedése az összvarianciából a rangsor függvényében.

Megbeszélés

Emlősök morfometriai vizsgálatánál a faktoranalizis főkomponens megoldása gyakran használt többváltozós módszer. Általános tapasztalat, hogy a legtöbb bélyeg szoros korrelációt mutat az első komponenssel s ezt ugy értelmezik, hogy az első komponens egy u.n. "méretarányos faktor". Az általam bemutatott példa esetében is hasonlót figyelhettünk meg, csak éppen az értelmezés nem ilyen egyszerü. Ugyanis, mint ahogy azt másutt közöltük [4], a nyers adatokat standardizáltuk, hogy az eltérő kor és földrajzi lokalitás méreteket befolyásoló hatását eltávolitsuk. Igy a méretarányos faktor relativ mértékének jelentős csökkenése lett volna várható, azonban ez nem következett be. Varimax rotációs főfaktor-eljárás sem eredményezett jobban értelmezhető eredményeket.

1. táblázat. Az első öt főkomponens faktorsúly mátrixa. A 0.25-nél kisebb értékek nem szerepelnek az első főkomponens szerint nagyságrendbe rendezett táblázatban. A változók röviditése Demeter és Lázár [4] szerint.

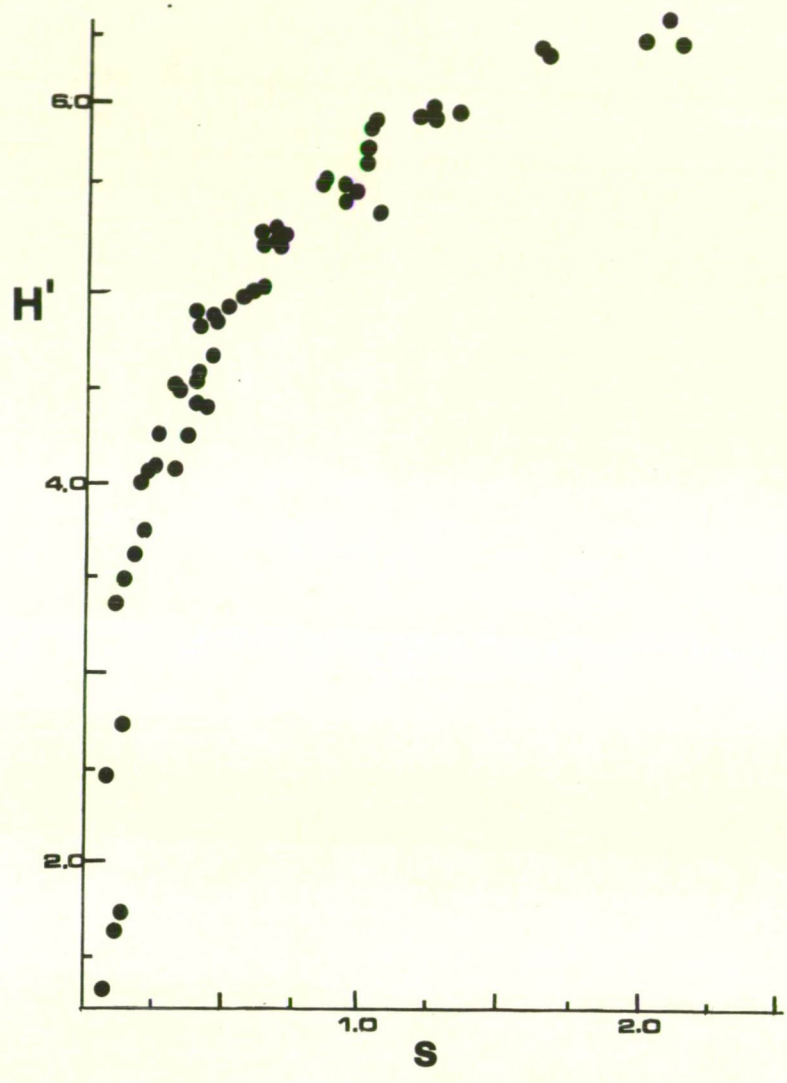
Változó	I	II	III	IV	V
PRPL	0.990				
DIM3	0.990				
GLES	0.990				
CBSL	0.988				
PRBS	0.987				
DAIC	0.987				
DACO	0.986				
BASL	0.981				
DAAN	0.979				
DBAB	0.973				
FNAL	0.973				
DICO	0.972				
LLMA	0.970				
LZYG	0.970				
HCON	0.969				
LDIAS	0.969				
DCOA	0.968				
HPAR	0.967				
LNAS	0.965				
WIRO	0.960				
WBRC	0.955				
HCOR	0.950				
DLEB	0.950				
WZGA	0.950				
LEPA	0.945				
EWFM	0.944				
WIIN	0.941				
HROS	0.940				
LHFT	0.939				
WIDA	0.936				
WIAB	0.932				
UMRC	0.930				
DBJP	0.927				
LMRC	0.921				
GHES	0.920				
BOWE	0.919				
HABL	0.917				
AHMA	0.914				
DBZP	0.910				
HCOK	0.902				
WINA	0.901				
LMRA	0.900				
UMRA	0.887				

1. táblázat folyt.

Változó	I	II	III	IV	V
WIM1	0.887				
HNAC	0.880				
HEAR	0.869				
LIOC	0.859				
HBRC	0.858				
LOTA	0.848				
WPM1	0.787				
DCOC	0.762				
SPSP	0.754				
WIFM	0.748	0.269			
HEFM	0.664	0.354			
WZYG	0.660		0.377		
DCNA	0.582		-0.313	-0.320	
LEFI	0.520				
WIPA		0.592	0.344	0.288	-0.370
VENZ	0.464	-0.507	0.427		
WICH		0.318	0.739		
VENX				0.771	0.288
VENY		-0.307		0.700	
DORZ		0.386			0.733
DORX	0.291		0.374		0.546
DORY	0.273	0.318	-0.433		

A diszperziós kritérium szerinti rangsorolás egyetlen változó jelentős szerepét mutatta ki, a módszer véleményem szerint alkalmas arra, hogy előzetes vizsgálatok alapján kiterjedt, most már kisebb számu változót figyelembe vevő morfometriai kutatás számára a legfontosabb bélyegeket kimutassa. Vegetációkutatásban széleskörben alkalmazzák ezt a módszert a vizsgálandó fajszám megállapitására [10].

Davies [2] az információtartalom szerinti rangsorolást értékesebb módszernek találta, mint a specifikus információtartalom szerinti karakterszelekciót. Ellentmondó eredményeink valószinü oka, hogy ő nominális skáláju adatokat vizsgált, mig az általam bemutatott példában intervallum skáláju adatok szerepeltek.



2. ábra. A bélyegek információtartalma az entrópia-függvény segitségével becsülve (H') a szórás függvényében

Irodalom

[1] Blackith, R.E. és Reyment, R.A. /1971/: Multivariate morphometrics. - Academic Press, New York.

[2] Davies, R.G. /1981/: Information theory and character selection in the numerical taxonomy of some male Diaspididae /Hemiptera: Coccoidea/. - Systematic Entomology 6:149-178.

[3] Davies, R.G. és Boratynski, K.L. /1979/: Character selection in relation to numerical taxonomy of some male Diaspididae /Homoptera: Coccoidea/. - Biological Journal of the Linnaean Society 12:95-165.

[4] Demeter, A. és Lázár, P. /1982/: Morphometrical analysis of the *Apodemus flavicollis - sylvaticus - microps* species complex in Hungary: a preliminary study of character selection and weighing. - Abstracts of Papers, Third International Theriological Congress, Helsinki /1982/: 54.

[5] Dixon, W.J. /ed./ /1981/: BMDP Biomedical Computer Programs. - Los Angeles, University of California Press.

[6] Kendall, M. /1975/: Multivariate Analysis. - Charles Griffin & Company Ltd, London.

[7] Orlóci, L. /1973/: Ranking characters by a dispersion criterion. - Nature, Lond. 244:371-373.

[8] Orlóci, L. /1975/: Measurement of redundancy in species collections. - Vegetatio 31:65-67.

[9] Orlóci, L. /1976/: Ranking species by an information criterion. - Journal of Ecology 64:417-419.

[10] Orlóci, L. /1978/: Multivariate Analysis in Vegetation Research. - Dr. W. Junk B.V., The Hague.

[11] Pielou, E.C. /1975/: Ecological Diversity. - Wiley, New York.

[12] Sokal, R.R, és Sneath, P.H.A. /1963/: Principles of Numerical Taxonomy. - Freeman and Company, San Francisco and London.